Calibración inteligente: una aplicación de la métrica

Nicolas Mantilla Molina[*]
Santiago Andrés Montes Camacho[**]
Universidad Industrial de Santander
Bucaramanga

5 de diciembre de 2022

Índice

Resumen

El estudio de la pureza del aire es fundamental en este momento en el que el cambio climático produce mayores impactos en la sociedad moderna, principalmente en terminos de la salud de la población. Evidentemente, una herramienta fundamental para ello es el seguimiento del material particulado en el ambiente, lo cual puede efectuarse por medio de sensores que garanticen precisión y fiabilidad en los datos obtenidos. En este orden de ideas, se lleva a cabo un proceso de calibración de un sensor de bajo costo tomando como referencia las mediciones obtenidas por un sensor de mayor calidad, en intervalos de tiempo similares. Para ello se efectúa un tratamiento de los datos que busca la determinación de una distancia euclídea, permitiendo así conocer su proximidad. De igual manera, se emplea una partición de los datos con el fin de establecer un conjunto de modelado y otro de prueba, que permitan la determinación de un modelo de calibración y la evaluación de su eficacia. De esta forma se obtuvo un mejoramiento del 46,7478 % en la proximidad de los datos, así como un comportamiento en la predicción de los datos de prueba que inviabilizaron la determinación de un alcance óptimo.

[*]e-mail: `nicolas2210707@correo.uis.edu.co`, código: 2210707
[**]e-mail: `santiiagoamontes@gmail.com`, código: 2210718

1. Introducción

El control de la concentración de los diferentes componentes en el ambiente es esencial en un contexto como el actual, en donde la contaminación dada por las diferentes emisiones diarias afecta de manera negativa trayendo consecuencias que repercuten a corto y largo plazo [1]. Es por ello que resulta importante llevar a cabo un seguimiento de estas concentraciones en diferentes puntos de una región, puesto que permiten conocer cómo se está comportando la pureza del aire a través del tiempo. Para ello, se cuenta con diferentes tecnologías como los sensores de detección de material particulado, los cuales facilitan el desarrollo de estos procesos.

En este sentido, estos instrumentos permiten realizar mediciones a través de mecanismos dados en su configuración, los cuales deben ser ajustados para obtener resultados confiables, por lo que es fundamental garantizar que el proceso que involucra la toma de datos se lleve a cabo de la manera más precisa, hecho que puede verse afectado por diversas causas (principalmente técnicas). En este punto, aparece una necesidad esencial: La calibración del instrumento, lo cual se emplea partiendo de una referencia que permite evaluar el nivel de certeza de la medición arrojada por el sensor en cuestión. En el presente estudio se efectuará un proceso similar, partiendo de los datos dados por sensores de medición de material particulado de bajo y alto costo en un intervalo de tiempo similar.

Teniendo esto en cuenta, en la sección 2 se hará una descripción detallada del proceso mediante el cual se realiza la calibración de el sensor de bajo costo. Seguidamente, en la sección 3 se abordarán los resultados obtenidos a través de un análisis del procedimiento empleado para la calibración. Finalmente, en la sección 4 se expondrán los principales hallazgos y aportes que resultan del estudio efectuado, en conjunto con algunas recomendaciones para futuras implementaciones y mejoras de la calibración.

2. Metodología

Para el análisis de la información, será fundamental establecer la manera de evaluar la proximidad de los datos dados por las estaciones IoT y las medidas de referencia de las estaciones AMB, en cuyo caso se manejará la distancia euclídea

$$\mathcal{D}(D_i, \hat{D}_{\hat{i}}) = \sqrt{\sum_{ik} (D_i - \hat{D}_{\hat{i}})^2}, \tag{1}$$

siendo D_i los datos de referencia y $\hat{D}_{\hat{i}}$ aquellos pertenecientes a las datos a calibrar. Debido a la necesidad de una correspondencia entre datos, se procederá a calcular un conjunto equivalente de datos mediante promedios móviles en los datos tomados. De acuerdo a esto, se define un tamaño para la ventana móvil y una frecuencia a la cual se realizarán los promedios, que implica un solapamiento de las ventanas. Cabe resaltar que la longitud de la ventana será determinada teniendo en cuenta el contexto físico de la situación, en este caso, el tiempo característico de cambio de la variable

climática $PM_{2,5}$.

Posteriormente, es oportuno llevar a cabo una partición de estos en el sentido que, con uno de ellos se lleva a cabo el procedimiento de regresión, y con el otro se evaluará que tan acertado resulta este ajuste para la calibración con un nuevo conjunto de datos, similar a las particiones de datos de prueba y entrenamiento empleados en *Machine Learning* para evitar fenómenos de sobreajuste o subajuste [2].

El proceso de calibración se realizará de la siguiente manera: se analizará la relación entre los datos de referencia y aquellos medidos por el sensor de bajo costo, efectuando sobre ello una regresión lineal cuya pendiente indicará qué tan similares resultan ambas mediciones, siendo ideal un valor esperado de pendiente $\alpha = 1$. Una vez obtenida la función de la regresión, será aplicada a los datos de prueba con el fin de evaluar qué tan acertada resulta la calibración. Dependiendo del comportamiento obtenido en este proceso, será determinado el alcance del modelo en cuestión partiendo de una tolerancia que será fijada según sea el caso.

3. El análisis y los resultados

Una vez cargados los datos brindados por los sensores de bajo y alto costo, fue considerada la información encontrada a partir de abril del 2019, debido a una gran discontinuidad de los datos presentada en intervalos de tiempo anteriores, lo que hace poco probable que se pueda extraer información útil del intervalo. Seguidamente, fue determinada una ventana móvil de una hora, tomando en consideración el contexto físico a partir del cual se está basando el sensor en cuestión: La determinación de material particulado que garantice una noción de sensibilidad frente a cambios dados en el tiempo, es decir, considerar como ventana intervalos de una hora resulta conveniente en el sentido que permite llevar un seguimiento más preciso, posibilitando el estudio del comportamiento de estas concentraciones en diferentes franjas horarias a través de los días.

Asimismo, fue llevado a cabo una comparación entre promedios móviles tomando saltos de 20 minutos y 1 hora, lo cual implica un solapamiento de 40 minutos (2/3 de hora) para el primer caso y una ausencia de solapamiento para el segundo. Con ello, se efectuó una regresión lineal para ambas situaciones, obteniendo un resultado como el mostrado en la figura 1, donde la recta es prácticamente la identidad. En este sentido, se optó por llevar a cabo un promedio móvil sin solapamientos, en relación a la eficiencia del cálculo en donde el aumento del tiempo de cómputo (más ventanas por promediar) realmente no brinda una mejora significativa en la relación de los datos.

Posteriormente, fueron evaluadas diferentes fracciones del conjunto total de datos, tomando en consideración la disminución porcentual de la distancia euclídea y las pendientes de regresión en la predicción (ver tabla 1). En este orden de ideas, se tomó una partición de 3/5 para los datos de entrenamiento, puesto que corresponde a la porción que cuenta con mayor mejoramiento en la distancia entre los datos predichos y de referencia, en cuya relación puede apreciarse una pendiente

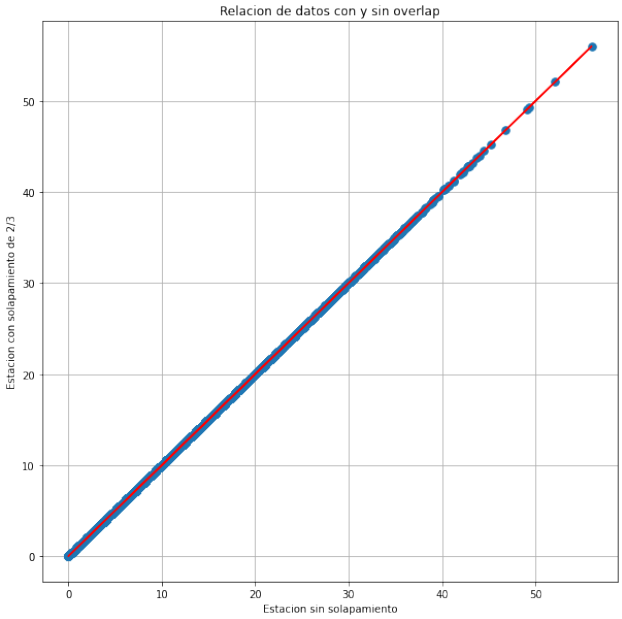


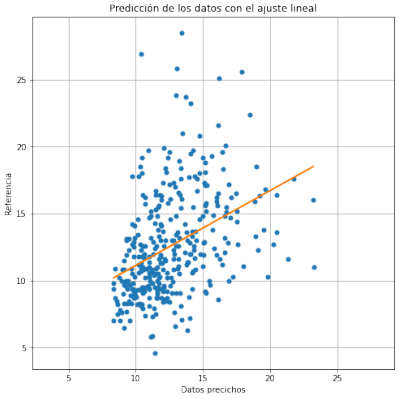
Figura 1: Regresiones lineales de la información obtenida por la estación de bajo costo sin solapamiento y con solapamiento de 40 minutos.

considerablemente próxima a la identidad $\alpha' = 1$.

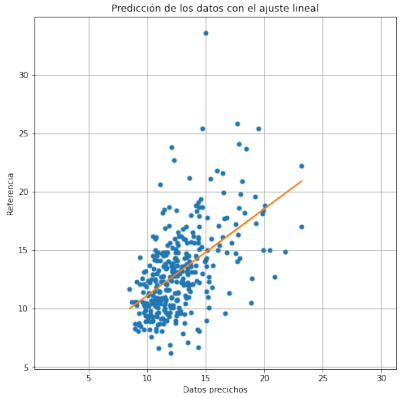
Porción utilizada	α	$d[\,\%]$	α'
1/3	0.4497	44.9816	0.8572
1/2	0.3802	45.2587	1.1197
3/5	0.3748	46.7478	1.1969
2/3	0.3609	46.0045	1.3394

Cuadro 1: Características de las regresiones efectuadas utilizando diferentes porciones de los datos para el modelamiento. α corresponde a la pendiente de la regresión, d es el porcentaje de reducción de la distancia euclídea (Eq 1) original, es decir: $\frac{|\mathcal{D}_0 - \mathcal{D}_i|}{\mathcal{D}_0} \times 100\,\%$, siendo $\mathcal{D}_i$ la distancia entre los datos calibrados y la referencia; y α' la pendiente de la recta que ajusta la relación entre estos últimos.

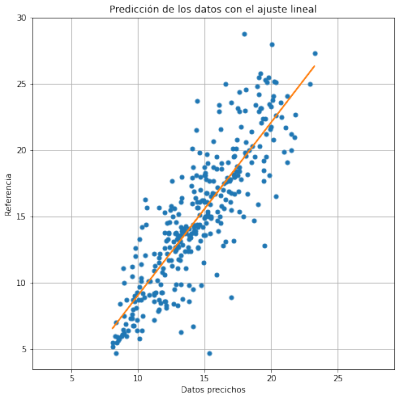
Así, fue llevado a cabo un seguimiento de la predicción del modelo a través del tiempo en el conjunto de datos de prueba, los cuales abarcan desde el 6 de julio del 2019 hasta el 31 de agosto del mismo año. Este monitoreo se empleó partiendo en cuatro particiones iguales de los datos, lo cual corresponde a lapsos de tiempo de aproximadamente 2 semanas, y se estableció la relación entre cada predicción y su referencia, tal y como puede visualizarse la figura 2.



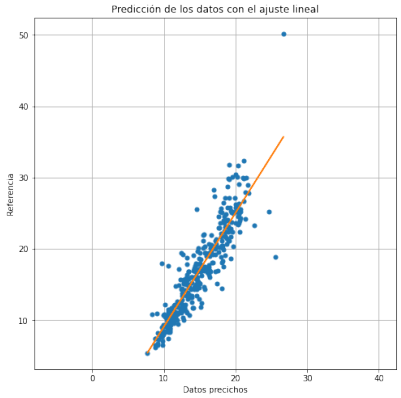
(a) Regresión lineal sobre la primera porción de los datos de prueba, $\alpha' = 0{,}558$.



(b) Regresión lineal sobre la segunda porción de los datos de prueba, $\alpha' = 0{,}742$.



(c) Regresión lineal sobre la tercera porción de los datos de prueba, $\alpha' = 1{,}3042$.



(d) Regresión lineal sobre la cuarta porción de los datos de prueba, $\alpha' = 1{,}5948$.

Figura 2: Regresiones lineales entre los datos predichos y los de referencia sobre cada 1/4 de los datos de entrenamiento.

A partir de ello, es posible apreciar una tendencia de aumento en la pendiente conforme pasa el tiempo, sin embargo, las mejores relaciones se obtienen en la segunda y tercera porción de la partición realizada, similar a las distancias euclídeas de cada una, las cuales resultan 70.753, 58.972, 55.392 y 74.243, respectivamente. Este comportamiento de los datos impide establecer un desmejoramiento directo temporal en la calibración, inviabilizando con ello la determinación de un alcance dado para el modelo empleado. Aún así es destacable la reducción de la distancia euclídea obtenida, siendo esta de un 53.26 % con respecto a los datos iniciales dados por el sensor, lo cual puede visualizarse a través de la figura 3, mediante una comparación entre los datos de referencia, los datos iniciales tomados por las estaciones IoT o el sensor de bajo costo y los datos dados por tal sensor tras su correspondiente calibración.

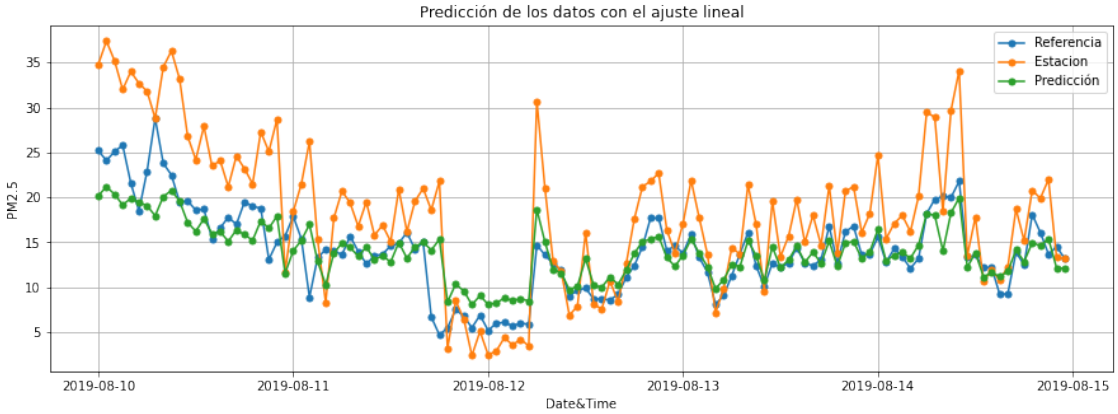


Figura 3: Concentraciones de material particulado dados por las mediciones de referencia, las mediciones tomadas por el sensor de bajo costo y las predicciones obtenidas mediante el modelo de calibración.

4. Conclusiones y Recomendaciones

Partiendo de los resultados obtenidos en el análisis, es posible determinar algunos aspectos importantes en cuanto al uso del promedio móvil como herramienta para el tratamiento de información, como lo es el hecho de no requerir un solapamiento entre las ventanas principalmente por el tamaño de ventana escogida, que coincide con la frecuencia con la que se tienen datos de referencia, así como datos de la estación de bajo costo. Si bien su funcionalidad podría intervenir en la suavización del comportamiento de la función, no implica realmente mayor diferencia, siendo por ende ineficiente para este caso.

Los resultados obtenidos en la predicción, pese a haber presentado una disminución en la distancia euclídea, demostraron un comportamiento contrario al aparentemente esperado, debido a que no se encuentra una desmejora progresiva conforme el paso del tiempo. Esto evidentemente dificulta la realización de procesos de determinación del alcance, lo cual es de gran importancia puesto que puede sugerir el rango de tiempo bajo el cual es oportuno llevar a cabo otra calibración. Una posible explicación a ello recae sobre la inviabilidad de emplear el mejoramiento total de la distancia en la predicción como factor determinante en la concreción de las particiones tomadas para este procesos. En otras palabras, sería recomendable replicarlo utilizando las demás particiones expuestas, y evaluando el comportamiento de la predicciones para cada distintas porciones dadas sobre el conjunto de datos de prueba.

Finalmente, se recomienda para futuras mejoras de la calibración inteligente utilizar diferentes modelos además del lineal, los cuales pueden describir de mejor forma la relación entre los sensores, además, se podría realizar una aproximación a la expresión que describe el material particulado en función del tiempo, mediante funciones periódicas, pues el comportamiento es aproximadamente periódico.

Referencias

[1] Clara Inés Pardo Martínez et al. Climate change in colombia: A study to evaluate trends and perspectives for achieving sustainable development from society. *International Journal of Climate Change Strategies and Management*, 10(4):632–652, 2018.

[2] Khan R. Z. Jabbar H. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). 2015.